Title

Method and Arrangement for Automatic Data Acquisition of Forms

Technical field

The present invention refers to a method and arrangement for the automatic data acquisition, by means of a means for the same, of forms whose design and information is not known in advance, by input into the said means together with storage of patterns of the same.

The prior art

It is a problem for companies, organisations and others to make good use of the information found in different types of paper forms, documents, etc.

With new, modern technology, these items can be scanned with a scanner and entered into a database via commercially available software programs. However, sorting, identification and other checking routines must to a large extent still be performed manually via the computer's display or screen.

For example, to store an invoice from one and the same company as one specifically designed document with a logotype and other visual elements, it must be revised so that its format is adapted to one that can be accepted by the software and then scored in a database. This and other procedures must be repeated each time an invoice with a new design is scanned with the software.

To identity an invoice from a company that is already registered, the whole invoice is often searched, which is time consuming. Certain software programs can have search routines that restrict the extent of this searching. It is, however, difficult to safeguard against blurred or hand-written lines of text, etc.

A need therefore exists for all who handle invoices and other forms to quickly be able to identify these and/or quickly be able to enter and store new patterns in their invoicing system.

Patent US-A-4 933 979 describes traditional data acquisition from forms and requires pre-defined templates/patterns with no self-learning (adaptive) ability.

Patent US-A-5 140 650 mentions data acquisition from forms with what is known as "Form out" technology to cover-over the original document and only retain the parts that are "filled-in." This data acquisition is often combined with data acquisition according to US-A-4 933 979. The patent does not have any adaptive function for data acquisition of unknown forms.

Another patent, US-A-5 293 429, concerns the classification of documents with the help of lines on the documents and does not directly concern data acquisition or any adaptive function

for this. USA-A-5 293 429 does not ensure the identification of lines with object areas (areas with text) and a "RCG-value" (ReCoGnition, a number that uniquely identifies a document).

None of the said patents generates a form map for a form unknown to the system according to the patent and stores the map in real time in a form database for recognition at the next opportunity for identification. For the inventions according to these patents, the unknown form must therefore be stored later by other means.

Summary of the invention

One of the objectives of the present invention is to solve the problems named above as well as others during what is known as automatic data acquisition (interpretation) in connection with the handling of paper-based information.

The present invention concerns a system (method and arrangement) for the automatic data acquisition of forms where the system has no prior knowledge of what the form looks like or where on the form the information is to be found. In this way, templates of forms do not have to be defined in advance, but are instead registered as they are submitted to the system, i.e., in real time.

To accomplish the above objectives, the present invention specifies a method and arrangement for the automatic data acquisition, by means of a means for the same, of forms whose design and information is not known in advance, by input into the said means together with storage of patterns of the same. The method is adaptive, by which it includes learning and registering of forms as patterns without filled-in text, and by it also including the following steps for accomplishing the adaptive registration:

generation of a form map based on the previously unknown form's design for identifying information contained on the form;

searching and comparing the form map with stored, registered maps in a means for storing form maps;

storing generated form maps in the storage means when they do not coincide with a stored map according to pre-determined limits for agreement;

indication of agreement according to the limits for agreement when agreement is found; and

continued data acquisition for identifying of the information content of the form.

According to one embodiment of the present invention, the form map can consist of an object

area list with objects contained in the form whereby the object comprises colours and/or wholly or partly of text.

In an alternative embodiment, the form map constitutes a line map comprising objects in the form of coloured lines from the form.

Horizontal lines in the line map are used to produce a horizontal key by dividing the form into a pre-determined number of horizontal segments along the y-axis in a cartographic system of co-ordinates, whereby each segment is equivalent to a position in the horizontal key.

Vertical lines in the line map are used to produce a vertical key by dividing the form into a pre-determined number of vertical segments along the x-axis in a cartographic system of co-ordinates, whereby each segment is equivalent to a position in the vertical key.

At least one line element that is included in a segment is marked in the equivalent key position, and segments that lack line elements remain unmarked in the equivalent key position.

A horizontal key and/or a vertical key constitute a line key in the line map, whereby during the said searching, the line key generated is compared with line keys stored in the means for verifying agreement.

The line keys are sorted in the storage means according to the number of markings.

The object's horizontal position in the object area list is used to generate a horizontal key by dividing the form into a pre-determined number of horizontal segments along the y-axis in a cartographic system of co-ordinates, whereby each segment is equivalent to a position in the horizontal key.

The object's vertical position in the object area list is used to produce a vertical key by dividing the form into a pre-determined number of vertical segments along the x-axis in a cartographic system of co-ordinates, whereby each segment is equivalent to a position in the vertical key.

At least one object that is included in a segment is marked in the equivalent key position, and segments that lack objects remain unmarked in the equivalent key position.

A horizontal key and/or a vertical key constitute an object key in the object area list, whereby during the said searching, the object key generated is compared with object keys stored in the means for verifying agreement.

The object keys are preferably sorted in the storage means according to the number of markings.

Searching results in a pre-defined number of requested probable candidates for the currently searched form.

If needed, an operator can support manually the whole or parts of the adaptive registration or identification of the new form or registered forms respectively if several alternative candidates are found as probabilities according to a factor of merit.

Finally, the identity of the form is confirmed by the data acquisition of a RCG-value.

Furthermore, the present invention specifies a arrangement for performing the above method.

The arrangement carries out automated data acquisition, by means of a means for the same, of forms whose design and information is not known in advance, by input into the said arrangement together with storage of patterns of the same. It learns adaptively and registers the design of the form, and includes a computer with the following means for carrying out the adaptive registration:

means for generation of a form map based on the previously unknown form's design for identifying information contained on the form;

means for searching and comparing the form map with stored, registered maps in a means for storing form maps;

means for storing generated form maps in the storage means when they do not coincide with a stored map according to pre-determined limits for agreement;

means for indicating agreement according to the limits for agreement when agreement is found; and

means for identification and continued data acquisition of the information content of the form.

In addition, the arrangement can include or constitute that specified according to the above method of the present invention, which is further illustrated in the accompanying non-independent claims for the arrangement.

Brief description of the drawings

Further reference to the enclosed figures and associated text will give a clearer understanding of the present invention.

Fig. I shows schematically how a line pattern is accomplished from a scanned-in invoice.

Fig. 2 shows schematically a flow-path for scanning, identifying, interpreting and validating a form according to the present invention.

Detailed description of preferred embodiments

In the continued description of the present invention, the forms are presented as invoices. However, the invention is not limited to invoice forms but also covers general documents containing text, figures, etc. as forms. Invoices are used here as an example of a form to exemplify the invention.

Fig. 1 illustrates schematically one part of invoice 10 that is scanned in a computer and that is shown on the display. As is evident from invoice 10, it is unclear or blurred after the scanning or input.

Invoice 10 consists partly of a logotype 12 and the vertical 14 and horizontal line 15 elements.

Note that even the logotype contains long black or varying degrees of shaded coloured line elements 16 that have been partly registered in a line map 18 according to the present invention, and that give an idea of what the original logotype 12 looked like, which simplifies identification when the invoice is an object to be identified as being as registered in a form map database. Coloured lines also include grey scales of black.

The form map that in this case constitutes line map 18 has been filtered from other objects 19, such as whole or parts of text objects or coloured objects, plus even the said line elements that include colour, which cannot be reproduced here, but that can be included as many coloured fields on a form 10.

An invoice 10 that is prepared according to the present invention, hereafter designated EH (Eyes & Hands), must be identified at an early stage. For successful identification, EH must have on a previous occasion, learned what the current invoice 10 looks like, which in simple terms means that information about the invoice is available in the form database in EH.

By necessity, the identification must be quick and be able to be made in a database that holds a very large number of invoice identities 18. It is not uncommon for databases to contain more than 10,000 identities 18.

The method and arrangement that EH uses does not require that an invoice is always put through a scanner in exactly the same way, i.e. the information on the invoice can vary somewhat in the x and y axes within a pre-determined measurement or threshold value. Fig. 1 shows a schematic cartographic system of co-ordinates.

In the present invention, (EH) comprises that in one embodiment, EH searches for all vertical 14 and horizontal line 15 elements of a pre-determined length on the invoice. Lines 14, 15 do not need to be free-standing and isolated, but can, for example, be part of a larger logotype text 12, such as ReadSoft AB in Fig. 1. The logotype 12 is represented as the line element 16 in line map 18.

The horizontal lines 15 and the vertical lines 14 constitute the basis for generating a horizontal key (h-key) and a vertical key (v-key) respectively according to the following:

* The invoice is divided into a large number of horizontal segments along the y-axis (not shown). Each segment is equivalent to one position in the h-key. If a certain segment includes one or more line elements 15, a mark or tag is placed in the equivalent key position. If not, an empty space, an inverted mark or anything else that differentiates itself from a mark is used instead.

* A v-key for the vertical line elements 14 is generated in a similar manner along the x-axis.

* The hand v-keys are given designations and together constitute a line key. Following this, a search is performed, which means that the current line key is compared with line keys for known invoices 10 that exist in EH's database. This comparison takes into account that individual lines or line elements 14, 15 can vary somewhat in position, plus that the total pattern of lines can be displaced somewhat according to suitable pre-determined values in the x and y directions, horizontally and vertically respectively.

* The line keys in the database are sorted according to the number of markings (tags), which are used to make the searching effective.

* The search results in a pre-determined number of probable candidates for the identity of the current invoice 10. All candidates are associated with a factor of merit or a probability that they are the current invoice 10.

* The identity of the invoice is finally confirmed by carrying out an interpretation of that known as the RCG-value (RCG- ReCoGnition). The RCG-value is a value at a given position that is unique for a certain invoice/supplier or other form. Examples of such values are bank giro numbers, post-office giro numbers, invoice numbers, total amounts, etc.

The said segments can, for example, form checked patterns that are fine-screened to varying extents according to the relative need for rapid searching.

The line keys can even be implemented on objects formed wholly or partly of text and colours. These are assigned line keys from an object area list that includes x and y-keys for the object. The object area list can, for example, consist of positions for certain selected objects. The

principles for line maps stated above are even appropriate for objects other than line elements to accomplish identification of forms.

If the line keys are not found in the database, this indicates that the invoice is not known, which results in the new line keys being stored in the database that, in this way, is updated in real time.

If necessary, the operator can, via his computer, manually support the whole or part of the adaptive registration and/or identification of a new form or registered form respectively if several alternative candidates are presented as probable according to the factor of merit.

In addition, the present invention includes an arrangement for performing the method according to the above.

The arrangement performs the automatic data acquisition, by means of a means for the same, of forms whose design and information content is not known in advance, by input into the said means together with storage of patterns of the same. It registers in an adaptive manner and learns the design of forms, and includes a computer with the following means for accomplishing the adaptive registration:

means for generating a form map based on the previously unknown form's design for identifying information contained on the form;

means for searching and comparing the form map with stored, recognised maps in a means for storing form maps;

means for storage of generated form maps in the storage means when they do not coincide with a stored map according to pre-determined limits for agreement;

means for indicating agreement according to the limits for agreement when agreement is found; and

means for identification and continued data acquisition of the information content of the form.

The, said means are preferably controlled by computer hardware and software, such as, for example:

A scanner for acquisition of data.

An electronic storage medium (hard disk, CD-ROM, etc.) for the means to store information

Signs, icons, signal generators, etc. for indicating purposes.

Filters and comparitors so that the means can search and compare, as well as filters and registers for identification.

On the whole, the means used in the present invention are well known to a skilled person in the technical field, but the way in which they are co-ordinated to achieve the object of the invention is, however, innovative.

In one embodiment of the present invention with reference to Fig. 2, a schematic flow-path is illustrated to show the scanning, identification, interpretation, and validation of a form according to the present invention.

Fig. 2 is divided by dotted lines into partial areas to clarify the different steps in a method according to the invention, whereby the steps constitute the scanning of the form 200, identification of the form 210, interpretation of the form 220, plus validating the form 230.

The form is scanned 200 into EH, and identification 210 follows. Identification consists of generating a line map 212, or alternatively an object area list, whereby an line key is generated. Following this, form 10 is compared 214 with known keys in the form map database, whereby a conformation of identification is obtained via the RCG-value. The next step includes deciding whether the identification was successful 216 according to the conditions "Yes" or "No". If the decision results in "No", a conditional investigation is made to sec if there are more candidates in the form of line keys 218. In the answer here is "Yes", a loop in the form of 214, 216 and 218 is performed until a successful identification is finally made, or until no further line key candidates are presented 218.

In the case of a successful identification, interpretation 220 of the form then begins by interpreting with the help of the current form map 222, after which validation 230 or evaluation 232 of the fields of the form 10 takes place. As an option, the operator can assist with selection if several alternative fields are found 234.

If the identification 210 is unsuccessful, and no further line keys are presented 218, interpretation 220 is performed in that self-learning with a form definition 224 is accomplished. The form definition consists of a template or a set of rules that describes the common elements of a specific collection of forms, for example, Swedish invoices. Following this, the RCG-value is interpreted 226 and a decision is made 228 whether the current RCG-value can be found in the form database. If the answer is "Yes", a re-interpretation begins 229, followed by a continued interpretation 222 that leads to validation 232.

If, on the other hand, the answer is "No", validation commences 230, 236, after which the form is saved in the form map database with the line key 238. Prior to steps 236, 238, the operator can, if several field alternatives are found, assist with the self-learning process.

The embodiments of the present invention described above are possible embodiments, but are not intended to limit the invention to such, as further embodiments will be evident to a skilled person in the technical area via the drafts of the enclosed claims.